Why we need Open Science and Open Education to bridge the corpus research–practice gap

Elen Le Foll

▶ **To cite this version:**

Elen Le Foll. Why we need Open Science and Open Education to bridge the corpus research–practice gap. Corpora for Language Learning: Bridging the Research-Practice Divide, pp.142-156, 2024, 9781003413301. 10.4324/9781003413301-11/need-open- . hal-04827282

HAL Id: hal-04827282
https://hal.science/hal-04827282v1

Submitted on 9 Dec 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

This is the author's accepted version (pre-copyedit manuscript) of the following chapter:

- Le Foll, Elen. 2024. Why we need Open Science and Open Education to bridge the corpus research–practice gap. In Peter Crosthwaite (ed.), *Corpora for Language Learning: Bridging the Research-Practice Divide*, 142–156. London: Routledge.

Please cite the version of record available from: https://www.taylorfrancis.com/chapters/edit/10.4324/9781003413301-11/need-open-science-open-education-bridge-corpus-research%E2%80%93practice-gap-elen-le-foll

Why we need Open Science and Open Education to bridge the corpus research–practice gap

Elen Le Foll, Department of Romance Studies, University of Cologne

https://orcid.org/0000-0002-5839-8010

Abstract

In this chapter, I argue that Open Science and Open Education practices are critical to addressing the much-discussed "gap" between applied corpus linguistic research and language learning and teaching practice. In particular, I focus on practical ways to facilitate the accessibility and sustainability of research findings and dissemination projects. Amongst other examples, I draw on my experience of leading a project in which pre-service English teachers co-created the Open Educational Resource (OER): "Creating Corpus-Informed Materials for the English as a Foreign Language Classroom: A step-by-step guide for teachers using online resources" (Le Foll, 2021; https://elenlefoll.pressbooks.com).

What do you mean by Open Science and Open Education and why do you think that these concepts are important in the context of using corpora for language learning?

Open Science and Open Education practices are all about ensuring that scientific knowledge is both rigorous and transparent, as well as accessible, collaborative, and inclusive. In practical terms, it is about sharing not only the results of academic research and teaching interventions, but also research data, testing instruments, methods, and educational resources openly – for the benefit of all.

In the context of language learning using corpora, we have seen time and time again that, although the concept of data-driven learning (DDL) has been around for nearly half-a-century now, there remains a wide gap between DDL research and what is actually happening in classrooms around the world. The number of academic publications on the potential of corpora for language learning appears to be ever-growing but, at the same time, uptake by teachers and learners remains very slow. There is no denying that they are many reasons for this, but I believe that one which has largely been ignored so far concerns the accessibility of

our research. The good news is that, unlike some of the other reasons, this is one that we can easily act upon.

What can those of us involved in corpus linguistics and DDL research do to make our research more accessible?

The reality today is that most academic research is hidden behind paywalls. These paywalls represent the most obvious barrier to the accessibility of our research for student teachers (especially those studying in the global South) and teaching practitioners worldwide. If we are serious about wanting to share knowledge, we should be prioritising non-profit open-access publication outlets and, when that is not feasible, uploading pre- and/or post-prints of our work on open-science repositories[1]. We should also consider exploring alternative forms of publications such as blog posts, podcasts, videos, and other online resources to communicate our research more effectively to non-academic audiences.

At the same time, if we want our research to contribute to generating cumulative knowledge about corpus linguistics for language teaching and learning, we must ensure that our methods are fully transparent so that others can reproduce our results (e.g., to check the robustness of published results using slightly different statistical methods), as well as replicate it in different contexts (e.g., in different countries, with learners of different proficiency levels and/or with different L1s and L2s) without having to always "reinvent the wheel"[2]. This entails sharing not just the results of our research but also the tools, methods, and materials used in any study.

The infrastructure to do so already exists. In addition to general research repositories such as the Open Science Framework (https://osf.io/) and Zenodo (https://zenodo.org/), there are a number of great initiatives for sharing and archiving linguistics research-related resources, for example: the IRIS database (https://www.iris-database.org/) and the Tromsø Repository of Language and Linguistics (https://trolling.uit.no/). As a discipline, I believe that our research could be both more efficient and effective if we engaged more in Open Science and Open Education practices.

How can Open Education practices contribute to making corpus linguistics more accessible for language teaching and learning?

Open Education is about teaching and learning with resources that are either in the public domain or licensed in such a way that everyone can engage in the so-called "5R activities". The five Rs are: retain, reuse, revise, remix, and redistribute. This means that Open Educational Resources (OERs) are freely accessible to all and can be adapted by anyone, for example by updating, expanding, or translating them. These new, modified versions can then also be shared with the community.

When we teach corpus literacy to (future) language teachers, it is important that we think carefully about the sustainability of what we teach: Can the activities we teach realistically be

[1] A very useful resource to find out more about the legal aspects of sharing pre- and post-prints of articles and monographs published with commercial publishers is: https://v2.sherpa.ac.uk/romeo/.
[2] I am not the first to use this expression in this context, see Braun (2007, p. 309).

used in the classroom? Will our students continue to have access to these corpus resources once the course/training is over? Can teachers readily adapt these corpus-based materials to their students' needs? We are lucky in that there are more and more open corpora and freely available corpus resources out there. But we need to make a conscious choice and effort to promote these in our teaching practice and, as researchers, contribute to them as much as we can.

Can you provide some examples of how corpus linguistics can be used to develop teaching materials that align with open educational principles?

The earliest example that I'm aware of is probably Tim Johns' collection of *Kibbitzers*, which were created in the late 1990s and are a remarkable example of an OER *avant la lettre*. These short DDL activities are available as PDF worksheets on https://lexically.net/TimJohns/. Each worksheet focuses on a real-life language question, mostly taken from English for Academic Purposes (EAP) classes. The materials feature selected concordance lines and provide guidance as to how to interpret these to solve each language issue. As such, they are an example of paper-based materials for DDL (see Boulton, 2010).

Since then, a number of projects have sought to integrate computer-based corpus literacy with language learning activities. Excellent recent initiatives that show how corpus linguistics can be used to develop teaching materials and that align with open educational principles include the *Corpus for Schools* platform and its associated teaching materials for English L1 and L2 (Gablasova et al., 2018), *BAWE Quicklinks* for English for Academic Purposes (Vincent & Nesi, 2018), the *Corpus-Aided Platform for Language Teachers* (Ma, 2018–) and the *Integrating Corpora* platform for German L2 (Vyatkina, 2020).

Together with some of my students, I also published an OER entitled: "Creating Corpus-Informed Materials for the English as a Foreign Language Classroom: A step-by-step guide for (trainee) teachers using online resources", available as an e-book on https://pressbooks.pub/elenlefoll/ (Le Foll, 2021). It not only provides ready-made corpus-based lesson plans, activities, and worksheets, but also explains to (pre-service) EFL teachers how to create their own corpus-informed materials. In total, there are 16 chapters divided into four sections providing lesson ideas for 1) Primary and lower secondary school, 2) Upper secondary school, 3) Content and Language Integrated Learning (CLIL) at secondary school, 4) Vocational education and English for Specific Purposes (ESP). This resource is one of the few that provides concrete examples of how to use corpora for language teaching and learning in pre-tertiary education in open access. Making this resource available as an OER aligns with Open Educational principles in that it allows anyone to not only access it but also to adapt it and update it, hence also contributing to the sustainability of the project.

Can you share some insights from your project on creating an OER together with trainee teachers? How did Open Science and Open Education principles guide your approach?

From the very beginning of the project, I decided to adopt an 'OER-enabled pedagogy' (Wiley & Hilton, 2018, p. 134). A key element of OER-enabled pedagogy is the concept of setting "renewable assignments" (Wiley, 2013). In contrast to 'disposable assignments' that primarily serve to assess whether students have met the course objectives and are frequently discarded

once they have been marked because they have no value beyond assessment, 'renewable assignments' not only support the learning of the individuals that produce them, but also result in new or improved OERs that that can benefit a wider community of learners. Examples of OER-enabled pedagogy projects include the creation and revision of anthologies, textbooks and online courses, and the creation and editing of Wikipedia articles, blogs and instructional videos (for many inspiring examples, see, e.g., Mays, 2017; Wiley & Hilton, 2018).

In this project, I decided that the renewable assignment would consist of writing a chapter for a (at first, hypothetical) OER textbook aimed at showing (trainee) teachers how to create their own corpus-informed teaching materials for the EFL classroom. The project took place at Osnabrück University (Germany) where I taught three iterations of a semester-long class on corpus linguistics for language teaching to M.Ed. students training to become EFL teachers for primary, secondary, and vocational education. Prior to starting to the course, students had little to no previous knowledge of corpus linguistics. Over the course of just one semester, they learnt about corpora, online corpus tools, and the benefits of using corpora for language learning. They then applied this newly acquired knowledge to develop their own corpus-based TEFL materials.

Although some students chose to write their chapters individually, they were strongly encouraged to co-write their chapters in groups of 2–4. Collaboration was a core aspect of the course: students gave each other feedback on all stages of their materials and chapter development. At the end of the semester, they wrote a term paper in the form of an OER textbook chapter, in which they presented their materials and how they could be integrated in a lesson plan, and explained to other (student) teachers exactly how they created these materials using online corpora and corpus tools (Le Foll, 2023). I made very clear that the eventual publication of their chapter in the OER was entirely optional, and that only original, high-quality contributions would be considered for publication.

In this project, OER-enabled pedagogy proved to be an effective framework to contribute to bridging the corpus research–practice gap in pre-service teacher training. In terms of its limitations, it's worth stressing that considerable time and energy was invested in the process of revising the chapters prior to publication. This happened during the semester breaks in both my and student authors' free time. An unforeseen limitation in terms of resources was that the online publication platform that I had chosen (https://pressbooks.com) changed its business model during the project. After this change, it was no longer possible to add or modify chapters within Pressbooks without paying for an additional subscription.

It's very difficult to empirically evaluate the success of such a resource, but from anecdotal feedback from colleagues, I know that the OER is being used in teacher training at many higher education institutions on all five continents. Looking back, it seems remarkable that, as a result of a one-semester course, corpus novices became proficient users of corpora and DDL multipliers with international reach!

In your opinion, what are the greatest challenges to creating and using Open Educational Resources about using corpora for language learning?

Before I discuss the challenges, I would like to stress that creating, using, and adapting OERs is a very positive and rewarding activity! In fact, after I published this OER that I co-wrote with my students, a number of colleagues launched similar initiatives – some of which are now online (e.g., Goulart & Veloso, 2023; Pinto et al., 2023). So, although there are challenges, these projects and the ones that I mentioned earlier are all proof that they can be overcome!

One challenge for hands-on corpus-based activities is certainly that there are still too few corpora that are suitable for language teaching and learning and that are freely available (or only for a limited number of queries, after registration, etc.). This is particularly true for languages other than English. Some are available for free, but their online interfaces are not particularly user-friendly, undergo frequent changes that make it difficult to create sustainable teaching resources, and/or contain language that are not always suitable for younger learners.

I was particularly keen to develop an OER on corpus-informed teaching materials together with trainee teachers because, in the process of working with corpora for the first time, my students came across challenges that I, as a corpus linguist, had not necessarily envisaged. The result is that, in their chapter contributions to the OER, the student authors describe, in simple terms and avoiding academic jargon, how they overcame these challenges. We hope that these explanations will prove useful to others – both (trainee) teachers, as well as academics who teach corpus literacy in pre- and in-service teacher training. As corpus linguists, we should also not underestimate the level of technical literacy required to use corpora for language teaching and learning. Even though today's student teachers and many practicing teachers may consider themselves "digital natives", this does not mean that they possess the digital and data literacy skills necessary to make use of corpus resources.

Finally, another limitation resides in the well-documented difficulty of making OERs known to the wider teaching community (for a recent survey, see Marín et al., 2022). In this endeavour, the academic community of Twitter was of great help. My posts on the OER project were widely retweeted by the #CorpusLinguistics and #TEFL communities. Sadly, however, only a small proportion of potential users can be reached in this way. I also presented the project at academic conferences, and it has been mentioned in several online seminars and courses. But, for all the positive resonance on these platforms, reaching non-academic audiences remains a challenge – even when applying Open Education principles.

What strategies can educators use to teach technical literacy skills to learners who may have limited experience with corpus linguistics and related technologies?

I personally think that it is important to start where learners are. To achieve this, educators can ask their students to solve some pertinent language issues and observe how they go about these tasks. Which resources are they aware of? Which ones do they use most frequently? In what order? How are they querying these resources? For a start, not all tasks are best solved with a corpus so, for some problems, they may already be using more appropriate resources. Next, we must make learners aware of the strengths and limitations of the resources that they are using (in my experience, most often Google, machine translation and, increasingly, ChatGPT and other AI-based tools) and, when meaningful, introduce them to corpora and corpus tools as an additional set of resources. Here, I believe that it is crucial to place emphasis

on using accessible corpus resources to solve concrete language issues that are of immediate interests to the learners.

In terms of teaching students the technicalities of using corpus tools, I have found that, in addition to going through practical examples in class, short tutorial videos that students can watch and re-watch at their own pace are very effective. For example, Laurence Anthony who develops the freeware corpus tool AntConc has an excellent series of video tutorials (https://youtu.be/_GSlwIO5QZE). Nowadays, creating such short video tutorials is quick and easy and, for those of us involved in teaching, sharing our tutorials in the spirit of Open Education is only a matter of few clicks. I must admit that when I uploaded the videos that I created for my students on how to use english-corpora.org and Sketch Engine (https://youtu.be/kDRpyJSE_6s), I did not envisage so many students and educators using them! Alternatively, renewable assignments in the spirit of OER-enabled pedagogy could revolve around students creating their own tutorials. Once these have been checked for quality, they can also serve as OERs (provided, of course, that the students consent to sharing their work with the wider community).

Another strategy that I rely on a lot when teaching the technical aspects of corpus literacy is the power of collaboration. Most of the tasks that I set my students are set up as group work. I believe that this is important because learning to use corpora for language learning and teaching is a complex set of skills. Hence, groups tend to fare better than individuals. Ultimately, I think that we view teaching corpus literacy as way to develop broader, transferable "life skills" that include digital and data literacy, basic statistical understanding, and critical thinking.

You spoke of the "technical aspects of corpus literacy". What are the other subcomponents of corpus literacy and how can educators go about teaching those following Open Education principles?

Callies (2016, p. 395) identified four subcomponents of corpus literacy: 1) understanding basic concepts in corpus linguistics, 2) searching corpora and analysing corpus data by means of corpus software tools, 3) interpreting corpus data, and 4) using corpus output to generate teaching material and activities. For both educators and learners, it is very tempting to focus on the second, technical subcomponent of corpus literacy. However, without a sound understanding of the basic concepts of corpus linguistics and without guidance and practice in interpreting the results of corpus queries, the potential of corpora for language teaching and learning cannot be unlocked.

That said, covering all four subcomponent of corpus literacy within more or less strict institutional constraints is no mean feat. Personally, I have found that OER-enabled pedagogy has been very beneficial in helping me to achieve this with my students. What is certain is that corpus tools and online corpus platforms will continue to change and evolve so that, if we want our teaching to be sustainable, we must ensure that we are teaching the principles and applications of corpus linguistics rather than "where to click".

How can Open Science and OERs promote more inclusive and equitable language teaching and research practices?

Using corpora, we can expose language learners to a much broader range of language varieties and registers than typically featured in textbooks and other commercially published teaching materials. For instance, there are more and more open corpora of varieties of spoken language, film and TV subtitles, and web registers that can help educators to explore situational language variation, something which is underrepresented in school textbooks and can create a genuine disconnect between curricular and extra-curricular language learning (Le Foll, 2022). We now also have access to some large open corpora that feature a range of language varieties that can help educators to broaden their teaching horizon to beyond the traditionally taught normative varieties. That said, whilst it is now much easier to create corpus-based materials and DDL activities to explore Cuban Spanish, Nigerian English or Swiss German, the number of languages for which this is possible currently remains very limited.

Open Science and Open Education practices promote accessibility, transparency, and collaboration, making scientific knowledge and educational resources available to all. As corpus linguists, we believe that these practices can empower language teachers and learners. Paradoxically, however, I have found that the use of corpora can provoke a lot of insecurities among (pre-service) language teachers. Many of my students expect and want to be able to deliver clear-cut answers to language questions, e.g., what is correct: *the kind of things* or *the kinds of things*?[3] For some, working with corpora is the first time that they are directly confronted with language variation. The realisation that correct-vs.-incorrect dichotomies on which they have so far relied do not hold in all contexts can be quite troubling. If we want teachers to adopt corpora in their teaching practice, we need to address this (see Le Foll, forthcoming). As I mentioned earlier, with often very little class time, it is tempting to focus on the technical aspects of corpora and corpus tools, but we must dedicate enough time to interpreting corpus results and discussing their implications for language teaching and learning.

Corpus linguistics can also be used to analyse and address social justice issues, including language-related bias and discrimination. For example, one of my students contributed an OER chapter demonstrating how to create a subcorpus of the News of the Web (NOW) corpus to create corpus-based materials for secondary level on the topic of Black Lives Matter (https://pressbooks.pub/elenlefoll/chapter/meise-reckefuss/).

How do you recommend that people get started with Open Science and Open Education?

I think that the process starts with a reflection on our personal motivation for doing research in this field and on the long-term impact of our research and teaching on language learning and teaching practices. To those keen to start implementing Open Science and Open

[3] For the curious, a quick search on *english-corpora.org/coca* reveals that the string *kind of things* occurs 1,380 times in the Corpus of Contemporary American English (COCA), whilst *kinds of things* is found 4,034 times. The combination of *kind of* + plural noun occurs 16,075 times, whilst a search for *kinds of* + plural noun returns 26,006 occurrences.

Education practices, I would say: Start by talking with your colleagues, then exchanging with the wider community; start by sharing pre- and post-prints of your papers, then move on to materials, code, data, and ideas. Some people seem to think that you either *do* or *don't do* Open Science and Open Education. I see them more as ideals to strive for where the journey is more important than the actual destination. And in keeping with that metaphor, it's worth remembering that even a journey of a thousand miles begins with a single step.

References

Boulton, A. (2010). Data-Driven Learning: Taking the Computer Out of the Equation: Data-Driven Learning. *Language Learning*, *60*(3), 534–572. https://doi.org/10.1111/j.1467-9922.2010.00566.x

Braun, S. (2007). Integrating corpus work into secondary education: From data-driven learning to needs-driven corpora. *ReCALL*, *19*(3), 307–328. https://doi.org/10.1017/S0958344007000535

Callies, M. (2016). Towards corpus literacy in foreign language teacher education: Using corpora to examine the variability of reporting verbs in English. In R. Kreyer, S. Schaub, & B. Güldenring (Eds.), *Angewandte Linguistik in Schule und Hochschule* (pp. 391–415). Peter Lang. https://doi.org/10.3726/978-3-653-05953-3

Gablasova, D., Brezina, V., McEnery, T., Meyerhoff, M., Reichelt, S., Arnold, T., & Cheung, C. (2018). *Corpus for Schools: Teaching English Language with Corpus Linguistics*. http://wp.lancs.ac.uk/corpusforschools/

Goulart, L., & Veloso, I. (Eds.). (2023). *Corpora in English Language Teaching*. https://pressbooks.pub/testbook123/

Le Foll, E. (Ed.). (2021). *Creating Corpus-Informed Materials for the English as a Foreign Language Classroom: A step-by-step guide for (trainee) teachers using online resources* (3rd ed.). https://pressbooks.pub/elenlefoll

Le Foll, E. (2022). *Textbook English: A Corpus-Based Analysis of the Language of EFL textbooks used in Secondary Schools in France, Germany and Spain* [PhD thesis, Osnabrück University]. https://doi.org/10.48693/278

Le Foll, E. (2023). 'Opening up' Corpus Linguistics: An Open Education Approach to Developing Corpus Literacy among Pre-Service Language Teachers. *Journal of Second Language Teacher Education*, *2*(2), 161–186. https://doi.org/10.1558/slte.25371

Le Foll, E. (forthcoming). *"To me, authenticity means credibility and correctness"*: A data-driven learning approach to encouraging pre-service teachers to re-evaluate their understanding of 'authentic English'. In C. Blume (Ed.), *Multiliteracies-aligned teaching and learning in digitally-mediated second language teacher education*. Routledge. https://hal.science/hal-04393791

Ma, Q. (2018). *The Corpus-Aided Platform for Language Teachers (CAP)*. https://corpus.eduhk.hk/cap/

Marín, V. I., Zawacki-Richter, O., Aydin, C. H., Bedenlier, S., Bond, M., Bozkurt, A., Conrad, D., Jung, I., Kondakci, Y., Prinsloo, P., Roberts, J., Veletsianos, G., Xiao, J., & Zhang, J. (2022). Faculty perceptions, awareness and use of open educational resources for teaching and learning in higher education: A cross-comparative analysis. *Research and Practice in Technology Enhanced Learning*, *17*(1), 11. https://doi.org/10.1186/s41039-022-00185-z

Mays, E. (Ed.). (2017). *A Guide to Making Open Textbooks with Students*. The Rebus Community for Open Textbook Creation. https://press.rebus.community/makingopentextbookswithstudents/

Pinto, P. T., Crosthwaite, P., Tavares de Carvalho, C., Spinelli, F., Serpa, T., Garcia, W., & Orenha Ottaiano, A. (2023). *Using Language Data to Learn About Language: A Teachers' Guide to Classroom Corpus Use*. The University of Queensland. https://doi.org/10.14264/3bbe92d

Vincent, B., & Nesi, H. (2018). The BAWE Quicklinks Project: A New DDL Resource for University Students. *Lidil. Revue de Linguistique et de Didactique Des Langues*, *58*, Article 58. https://doi.org/10.4000/lidil.5306

Vyatkina, N. (2020). *Incorporating corpora: Using corpora to teach German to English-speaking learners [Online instructional materials]*. Retrieved from https://corpora.ku.edu

Wiley, D. (2013). What is Open Pedagogy? [Open content]. *Improving Learning*. https://opencontent.org/blog/archives/2975

Wiley, D., & Hilton, J. L. (2018). Defining OER-Enabled Pedagogy. *The International Review of Research in Open and Distributed Learning*, *19*(4). https://doi.org/10.19173/irrodl.v19i4.3601